DOCUMENT RESUME

ED 274 381 JC 860 485

AUTHOR Hild, Harold N., Ed.
TITLE Developmental Learning: Evaluation and Assessment. NARDSPE Research Report, Number 1. First Edition.
INSTITUTION National Association for Remedial and Developmental Studies in Post-Secondary Education.
PUB DATE 82
NOTE 56p.; Table contains small type.
AVAILABLE FROM National Association for Developmental Education, P. O. Box 60227, Chicago, IL 60660 ($6.00).
PUB TYPE Reports - Descriptive (141) -- Viewpoints (120) -- Collected Works - Serials (022)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Criterion Referenced Tests; *Developmental Studies Programs; Educational Diagnosis; Educational Testing; *Evaluation Methods; *Naturalistic Observation; Postsecondary Education; *Program Effectiveness; Remedial Programs; *Standardized Tests; Student Evaluation; Student Placement

ABSTRACT

Three types of evaluation methods are discussed in terms of their use in evaluating both cognitive and affective learning in developmental programs: standardized tests; criterion-referenced tests; and naturalistic inquiries. Section I looks at the complexities of learning and the inherent difficulties in judging the quality and quantity of what is learned by students in developmental programs, and proposes a model that can be applied to the evaluation of learning in postsecondary developmental programs. Section II describes three general categories of tests (i.e., placement, diagnostic, and achievement) and specific types of tests (i.e., standardized, criterion-referenced, habit and attitude inventories, and teacher-made competency tests). These tests are analyzed according to their strengths, weaknesses, and uses. Additionally, section II reviews several instruments that can be used in developmental programs, and section III offers a rationale for the utilization of criterion-referenced tests in developmental programs and offers suggestions for developing these tests. Section IV shows how unobtrusive or naturalistic methods of evaluation can be used to fill in the gaps left by other tests, including a definition of naturalistic inquiry and a description of its basic principles. Finally, a concluding section highlights major points and implications for the future of evaluation of developmental studies. (EJV)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☒ Minor changes have been made to improve reproduction quality.

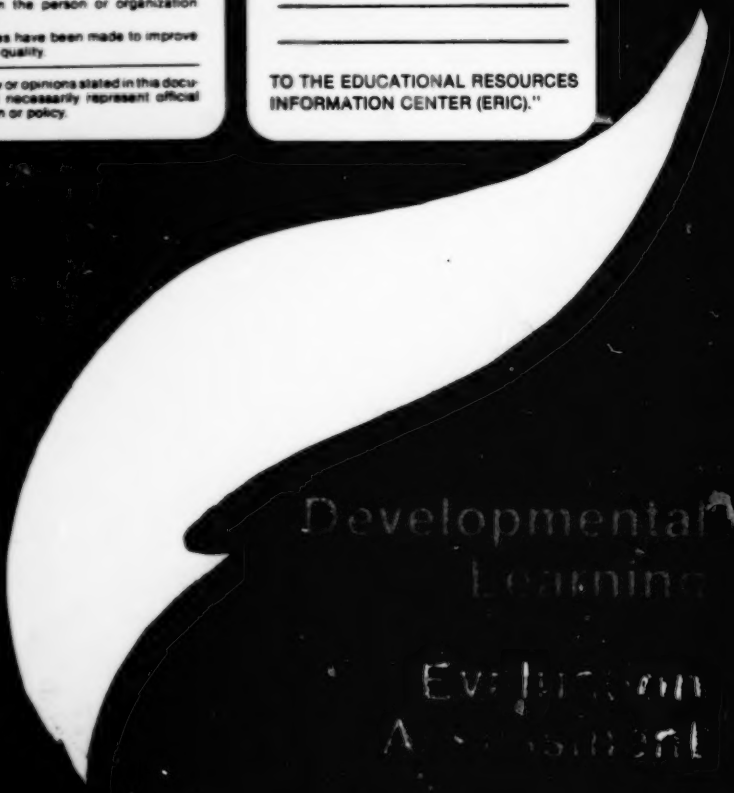
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

A. Reynolds

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED274381



Developmental
Learning
Evaluation
Assessment

JC 860 485

NARDSP

Developmental Learning:
Evaluation and Assessment

Edited by:
Harold N. Hild

Written by:
Carol Clymer Walvekar
Nancyanne Rabianski
Karen G. Smith

A Research Report Prepared for the National Association for Remedial-Developmental Studies in Post-Secondary Education.

Spring 1982

Development Learning: Evaluation and Assessment edited by Harold N. Hild. Copyright © 1982:

The National Association for Remedial and Developmental Studies in Post Secondary Education

First Edition 1982.

FOREWORD

Five years ago a small group of enthusiastic educators decided to form an association which they felt would help them achieve several short and long term goals. That small group meeting in Chicago has seen many of these goals realized over the past several years: an annual conference attracting major practitioners from across the nation; a newsletter published regularly; increased membership; the development of regional and local chapters; and an executive committee capable of providing effective leadership and needed resources to the field. With the publication of this research report, another major goal has been attained.

I wish to take this opportunity to acknowledge and thank those who have made the attainment of this goal as well as those aforementioned possible. In the forefront of any such acknowledgment must lie those sturdy and stalwart charter members from that first year. Equally in that forefront are Chicago State University and Northeastern Illinois University, whose financial and moral support in those early years was essential. Similarly, those early executive officers, especially Gary Saretsky, deserve a heartfelt handshake of appreciation and thanks.

More close in time, the special efforts of Carol Clymer Walvekar, Nancyanne Rabianski, and Karen Smith in the coordination and writing of this report should be applauded and recognized. Behind the scenes the continuing support of the present executive board, especially Hunter Boylan, has provided the necessary environment for completion of this research project. Lastly, Audrey Reynolds, Newsletter Editor, whose proofreading and technical expertise have helped ensure a quality publication.

In conclusion, I would like to dedicate this first publication to the membership of the National Association and express an earnest hope that others in this series are soon forthcoming.

Harold N. Hild
Editor

CONTENTS

INTRODUCTION ..... Page 1
By Carol Clymer Walvekar

SECTION I ..... Page 7
Evaluation of Learning: A Model for Developmental Education
By Carol Clymer Walvekar

SECTION II ..... Page 19
Testing in Developmental Programs
By Nancyanne Rabianski

SECTION III ..... Page 35
Developing and Using Criterion-Referenced Measurements in Developmental Programs
By Karen G. Smith

SECTION IV ..... Page 35
Using Naturalistic Methodologies to Evaluate Learning in Developmental Programs
By Carol Clymer Walvekar

SUMMARY ..... Page 59
By Carol Clymer Walvekar

REFERENCES ..... Page 63

ABOUT THE AUTHORS ..... Page 67

INTRODUCTION

By Carol Clymer Walvekar

The word *learn* is used daily by educators. We think we know what *learn* means; rarely do we ponder over its definition as it is a concept common to us. The definition supplied by *Webster's Seventh New Collegiate Dictionary* is on the surface quite simple: "to gain knowledge or understanding of or skill in by study, instruction, or experience." However, when we closely and separately examine each of the key words included in this definition, we find that the meaning of *learn* is considerably more complex. For example, knowledge is the collection of information, ideas, and concepts that are known to an individual, society, or field of study. Yet, "knowledge", in and of itself, is hardly tangible and it is constantly changing. "Understanding" is another intangible term which varies greatly from individual to individual. One person's understanding of a concept is frequently not only different from the next person's but can change as an individual matures or as knowledge changes. Observe, for instance, the concept of nuclear power: some "understand" that nuclear power is a very viable solution to many of our energy problems, while others "understand" that it is a dangerous alternative that could prove fatal to the human race. Interestingly, each of these two opposing views is based on a substantial amount of knowledge. The words "skills," "instruction" and "study" are more concrete, but the range of meanings and connotations of each of these terms is extensive. Finally, as no two people have the same "experience," this component of learning includes an infinite number of possible definitions. Learning and all of its components are complex entities. When we attempt to define the entire concept, not to mention measure or evaluate it, precision and consensus are difficult to achieve.

It is erroneous to suggest that we have not been trying to develop a truly functional definition of both learning and the evaluation of it. Learning and its evaluation have been concerns of educators for many centuries. Eblinghaus, Pavlov, Watson, Lashley, and Skinner laid the groundwork for modern learning theory (Marx and Bunch, 1977), while Golton, Gattell, Thorndike, Binet, and Spearman (Lien, 1976) helped to lead the way for the theory of measurement of learning. While the purpose of this booklet is not to present and defend a functional definition of learning[1], it is important to acknowledge the difficulty that exists in defining it. Bruffee (n.d. p 4-6) speculates on the difficulties involved:

"What we understand knowledge to be, what we think it

[1]Many definitions of learning exist and the reader is directed to Hilgard and Bower (1966) and Marx and Bunch (1977) for more detailed accounts.

> means to know something, and hence what we think it means to learn it, have changed radically during the first half of this century...until the late eighteenth century, the supreme point of reference was an almighty God... From the eighteenth century on, the fixed point of reference became almighty math.. In short, before the twentieth century most educated people believed, and many still do, that the value and certainty of anyone's claim to know something is a function of its proximity—that is, its structural similarity—to knowledge at the top of the pile: mathematically validated knowledge...and we disseminate our scientistic faith by our reliance, throughout the curriculum of American secondary and post-secondary education, on 'objective testing."

Bruffee is suggesting that the "facts" we use to determine learning have been dictated by mathematically based knowledge and understanding. This reliance on mathematics has, in my opinion and Bruffee's, caused a great many learning theorists to view learning and the evaluation of it in terms of changes in behavior and performance which are observable and quantifiable. This view has resulted in the development of thousands of tests which are designed to measure the extent of an individual's knowledge, skill or understanding. In short, we need to have "numbers" as evidence of what is learned. Numbers do not lie; they are stable and predictable. Two plus two is always four. Therefore, we naturally look to mathematics to support our judgments and decisions about learning, but this is not always justifiable because often learning takes place which is not easily quantifiable or measurable. For example, we may not be able to use tests to measure an individual's reasoning ability, appreciation of a subject, inventiveness, or creativity. Naturalistic inquiry, an area of research using such methods as controlled observation, interviews and surveys, will help us to evaluate the affective areas of learning which are not easy to quantify.

The sections that follow will discuss three types of evaluation methods that can be used to evaluate both cognitive and affective learning in developmental programs: standardized tests, criterion-referenced tests and naturalistic inquiries. Section I sets the stage for the evaluation of learning in developmental education by providing a model that can be applied to developmental programs in post-secondary institutions. Section II describes three general categories of tests: placement, diagnostic and achievement. It also discusses specific types of tests: standardized, criterion referenced, habit and attitude inventories, and teacher made competency tests. These tests are analyzed according to their strengths, weaknesses and uses. Additionally, Section II reviews several instruments that can be used in

developmental programs. Section III offers a rationale for the utilization of criterion-referenced tests in developmental programs and gives suggestions for developing them. Finally, Section IV details unobtrusive measures or naturalistic methods for evaluating learning. Finally, the conclusion of this booklet contains highlights of major points and implications for the future of evaluation of learning in developmental programs.

SECTION I: EVALUATION OF LEARNING: A MODEL FOR DEVELOPMENTAL EDUCATION

By Carol Clymer Walvekar

Unfortunately, evaluating the learning skills of students in remedial and developmental programs is not an easy task because of the indeterminate nature of the many entities which cause, influence, and constitute learning. It is fairly safe to say that most people who are reading this publication realize that we have many students in post-secondary institutions who enter without adequate reading, writing, math or study skills for learning in college. Consequently, the institution of remedial and developmental programs is common in higher education. It is also fairly safe to say that the success of remedial and developmental programs is highly dependent on a program's ability to continually determine the extent and nature of each participant's learning skills. Furthermore, the demonstration of retention of learning skills for remedial and developmental students, once they have matriculated into other curricula, is another more complicated factor that contributes to the success of the program. Difficult as it may be to supply accurate information about learning and the retention of it, we must. And, we must do an effective job of it or our students, schools and programs will suffer. To begin to fulfill this somewhat formidable responsibility, we must have a clear understanding of the purposes of evaluating learning. The purposes can be very broadly categorized into six general areas. These areas have been identified by Ball (n.d. 37,40).

1. To contribute to decisions about program installations.
2. To contribute to decisions about program continuation, expansion (or contraction), and/or "accreditation".
3. To contribute to decisions about program modification.
4. To obtain evidence favoring a program to rally support.
5. To obtain evidence against a program to rally opposition.
6. To contribute to the understanding of basic processes.

Although these purposes were generated for the context of overall program evaluation and planning, they can be easily applied to the evaluation of an individual's learning in developmental programs. For example, evaluation related to Purpose #1, "decisions about program installation" would be made through the use of diagnostic and placement testing and student information questionnaires completed on the first day of class or during an initial interview. Information gained from these evaluation tools would help the staff to initiate the appropriate instructional program. Next, competency testing, criterion-referenced testing, teacher made paper/pencil tests, or observations made during interviews are examples of evaluation methodologies that would be used for making "decisions about

program continuation, expansion (or contraction), and/or accreditation," purpose #2. For example, if a commercially made instructional program is utilized in a developmental course, the instructor must continually determine if the instructional program prescribed is meeting the learner's needs. If students are not grasping a particular concept, certain modules of the program may have to be repeated or supplemented with other materials. Other modules might be omitted if they seem irrelevant to the curriculum of the particular institution. Post tests, unit tests, or remarks from other faculty may indicate to the developmental program's administration that too many components of the commercially prepared program are not preparing developmental students for non-developmental courses. In such cases, developmental program administrators would use test results and feedback from faculty to "modify" the program so that it meets the needs of the institution, and thereby conduct an evaluation for purpose #3, "program modification." Certain administrators may want to eliminate a commercially prepared instructional program used by the developmental division because it is expensive to maintain, but perhaps instructors feel that it indeed does prepare students for transfer into other divisions. Attrition data on students who have successfully used the instructional program might be compared to attrition data of students who have not used the program to show "evidence...to rally support", purpose #4. Conversely, if the data from an evaluation reveal that the program is actually detrimental to the students and the institution, then evaluation results would be used to try to convince advocates of the program that it should be eliminated. Hence, the purpose of the evaluation here is to gather "evidence against a program," item #5 on Ball's list. Finally cognitive mapping could be completed on all students using a developmental program and data on the cognitive pattern of students could be analyzed and compared to grades to "contribute to the understanding of basic processes," the last purpose for evaluation, according to Ball.

Ball's broadly stated purposes provide a very general framework for organizing systems for evaluating learning; however, a truly practical delineation of purposes is much more specific. These practical purposes range from trying to determine the areas of phonics that a student must learn in order to improve spelling skills to the effect that the testing environment has on students' test scores. Both kinds of purposes must be clearly defined.

Once purposes for the evaluation have been identified, the evaluator must decide on a system for evaluating the learning of the developmental student. This system should function on an ongoing basis from the time of entry to the program to the time of exit from the program. Below, an example of a system is described.

A college student applies to a postsecondary institution, and on the

basis of high school transcripts, the admissions office recommends enrollment in a developmental reading and study skills program. The student is required to participate in the following evaluation activities: On the first day of classes, the student reports to the testing center and takes several tests: The Nelson-Denny Reading Test, the Canfield Learning Styles Inventory and the McGraw-Hill Test of Basic Writing Skills. The student is also asked to complete a questionnaire that solicits information regarding the individual's high school learning experiences, reading habits, special difficulties with learning particular subjects, other scheduled classes and instructors, interests, reasons for enrolling in school, and the reasons for taking the reading and study skills course. Finally, the student must write a short paragraph that describes his or her educational goals. One day later, after returning to the developmental class, the student receives a diagnostic sheet with an assessment of strengths and weaknesses in reading, writing, and study skills. Also, the student is given an Individualized Educational Plan (IEP) that identifies the instructional objectives and materials that should be completed in class. Written on the bottom of the plan are the words, "Please make an appointment to meet with your instructor as soon as possible." Upon conferencing with the instructor, the student learns that more diagnostic tests must be taken because the other tests did not exactly pinpoint all of the areas of difficulty. On the following day, the student takes the Wide Range Achievement Test/Spelling and an informal reading inventory. After the test results have been analyzed, the student is appraised of more specific learning difficulties and some interim objectives and more materials are added to the IEP. During the developmental class, the student learns about taking notes, studying textbooks, using the library, writing research papers, studying for exams, and managing time. After every class, an assignment is required that covers the material the teacher presented in class. Once a week, a test is given on the material covered in class. In addition to class, the student must attend reading and writing labs and work in the various instructional materials written on the IEP. Once a week, a meeting with a tutor is arranged and the student receives individual instruction to review and supplement the information that is learned in the developmental class and reading and writing lab. Throughout the sessions, the tutor records observations about the student's progress. Additionally, the student maintains a notebook with the work from assignments, questions about the materials or exercises, and any difficulties that are apparent while working on the assigned materials. Occasionally, the notebook contains notes about boring or stupid exercises, or trouble with concentration because of noise or distractions. Periodically, the tutor administers verbal and written tests on the material that is being learned by the student.

At the end of the year, the student goes to the testing center to take similar tests as those that were completed at the beginning of the semester. They do not have the same stories or questions as before, but the format is exactly the same. In a final conference with the student, the instructor conveys that the student's work has been satisfactory because objectives were accomplished and test scores improved. However, the student is informed that the second level of the developmental course must be taken before his/her reading and writing skills will be adequate for college work. In the next developmental class, the student repeats essentially the same process only using more advanced materials.

The system for evaluating learning that has just been described is probably familiar to most developmental programs practitioners. A model of this system can be seen in Figure 1.

This model is based on a theory of evaluating instruction developed by M. C. Wittrock (1969). He contends that the evaluation of instruction is characterized by the cause and effect relationship of instruction and learning.

Specifically, Wittrock (p. 3) defines the *evaluation of learning* as making "explicit the changes in students' behavior to try to determine what had been learned during instruction." He sees the *evaluation of learners* as making "explicit the students' abilities, interests, and achievements to determine student performance at the end of instruction." *The evaluation of learning environments* involves examining such areas as time, materials and their sequencing, the learning facility, staff training, and so forth. Evaluating the learning environment and such areas as the effectiveness of the staff, the adequacy of materials, and the amount of time alloted to learning are topics which require thorough explanation and perhaps another publication altogether. With this notation made, I shall describe specific evaluation procedures and methodologies that should be used in order to utilize the model in developmental programs.

As I have already stated, in a very broad sense, evaluation is conducted for six basic purposes all of which relate to decision making. In order to organize evaluation so that we can make the necessary decisions, the practitioner should try to answer four simple questions:

1. What is it that I am evaluating?
2. How am I going to evaluate it?
3. When am I going to do it?
4. Why am I doing it?

The *What* of evaluation is evaluating the learner, the learning environment and eventually the learning that takes place, and it is accomplished in three stages. The first stage occurs *when* the student enters the program. During evaluation, the practitioner must consider both the affective and cognitive domains of the learner as both

BEST COPY AVAILABLE

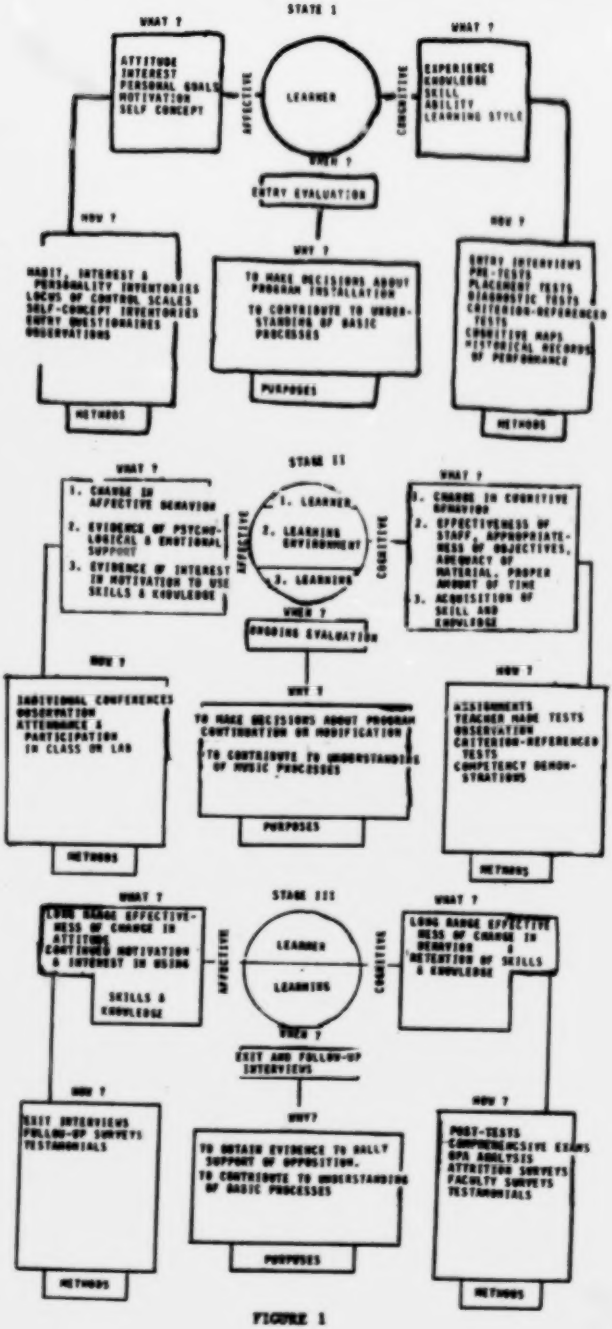


FIGURE 1

domains greatly influence learning outcomes. On the affective side of the model, the attitudes, interests, goals, motivation, and self-concept of the learner are represented. Some methods used to evaluate these affective areas are habit, attitude and personality inventories; self-concept and locus of control scales; entry questionnaires and interviews; and notations about observations of affective behaviors that are exhibited by the learner.

On the cognitive side of the model, assessment of the learner's experience, knowledge, skill, ability and learning style is attempted. This is usually accomplished by administering a variety of tests, mapping cognitive styles and analyzing the historical records of learners. Essentially, I have just described *how* to evaluate the learner, but it should be noted that a practitioner would not necessarily use all of the methods described to determine the learner's level when entering a program. The entry interview should be the controlling method of evaluating the learner during Stage I, so that information obtained in the interview can be used to decide which tests and inventories should be administered. Moreover, it is possible for an entry interview to reveal enough information to enable the evaluator to make decisions about the instructional objectives of the learner, without giving any diagnostic tests.

Stage II provides the data for determining if the long-range objectives need to be broken down into short-term or interim objectives. Thus, the information obtained in the second stage of the model enables the evaluator to make decisions about modifying the instructional program to meet the learner's needs. Here, when we evaluate, we not only assess 1) the learner, 2) the learning environment, and 3) the learning, but the interactions of all three entities as well. Consequently, the evaluation process in Stage II can become very complicated because of these interactions. The evaluator must be careful to remember that there is a cause and effect relationship between all three entities: the existing characteristics and abilities of the learner effect the conditions of the learning environment which together effect the learning.

Additionally, we must again consider both the affective and cognitive influences on the learner, the learning environment, and the learning. In evaluating the learner during the learning process we are trying to discover if there are cognitive and affective changes in behavior. In order to ascertain if a change in behavior has taken place, we look for the learner's acquisition and application of skills and knowledge, and for evidence that the learner is interested and motivated to use these. Of course, the learner must have an appropriate learning environment: an effective instructor who is well trained and capable of teaching what needs to be learned and adapting to the needs of the students, appropriate instructional objectives that are

geared to the needs and abilities of the learner, adequate materials that are at the instructional level of the learner and that are interesting and properly sequenced, and the proper amount of time for meeting the instructional objectives. Finally, there must be evidence that the environment in which the learning takes place is psychologically and emotionally conducive to learning. For example, an instructor might take much time to write an IEP that exactly addresses the needs of the learner, but explain it in a condescending tone so that the learner lacks confidence in his/her ability to work through the plan.

As mentioned before, a discussion of the methods of evaluating the learning environment requires another publication in that it is really program evaluation and another subject altogether. The reader is directed to other sources for more information on this subject.[1]

Evaluating the learner and *what* he or she is learning during the time of instruction occurs as an ongoing process. Homework or in-class lab assignments, tests made by the instructor, criterion-referenced tests, demonstrations of competency and observation can and should all be used constantly during instruction to determine the amount and quality of the cognitive learning that takes place for the learner. Evaluating affective learning is more difficult because it is not that easily measured. One way to evaluate changes in affective behavior is through observation. Much data can be collected by watching the learner work, listening to comments made during conferences, and recording the observations. Participation and attendance in class, to some extent reveal the amount of interest and motivation that a learner has, but it is certainly not conclusive evidence of the extent of motivation. Some individuals may be very motivated to learn, but shyness or insecurity keep them from participating fully in class. Additionally, some learners have poor attendance records if the instructor and materials are boring or incompetent. (For a more detailed description of utilizing observation as an evaluation tool, see Section IV of this publication.

Stage III of this model portrays what we usually do at the end of instruction. At this point, we want to know what the learner has learned. We want to know if the learner has improved in weak areas, increased skills, accomplished objectives, and met needs. And eventually we want to know that the skills and knowledge have been retained and that the learner is actually motivated and interested in using them. At the end of the instruction, we administer post-tests

[1]Dressel, P. L. Handbook of Academic Evaluation, San Francisco: Jossey-Bass, 1978.
The Joint Committee on Standards for Educational Evaluation. *Standards for Evaluations of Educational Programs, Projects, and Materials.* New York: McGraw Hill, 1981.
Morris, L. L. (edt.) *Program Evaluation Kit.* Beverly Hills: Sage, 1978.

and/or comprehensive exams to assess the skills and knowledge of the learner. Also, an exit interview can reveal positive or negative changes in the learner's self-concept and self-confidence in using the learning skills which have been developed. Support for observations made during the interview may be secured by administering habit, attitude, interest, and personality inventories.

After the learner is no longer a participant of the developmental program and has moved into other parts of the curriculum, we want to know if the learner is retaining and using the skills and knowledge that were evident at the time of exit from the program. It is difficult to find out the long-range effectiveness of changes in cognitive and affective behavior which occur as a result of the learner's participation in the developmental program as there are so many variables which influence learning and the learning environment in the general curriculum that cannot be controlled. However, if we analyze the GPA and attrition rates of students who complete the developmental program, we can at least determine patterns. If developmental students have higher GPA's and lower attrition from other courses than non-developmental students, we cannot conclusively say that the difference is a result of the developmental program. We can, however, note that there *is* a difference. On the other hand, if the developmental students have lower GPA's and higher attrition rates than non-developmental students, we have an indication that the developmental program is not doing what we want it to do. Again, in this case, we do not have conclusive evidence, but we do have data that signal problems somewhere in the developmental program.

We can also acquire follow-up information on the effectiveness of the cognitive and affective learning by conducting surveys. A good way to determine if the skills and knowledge have been retained is to survey the faculty who teach the developmental students in the degree programs of the institution. Examples of questions that should be asked are:

1. Are the developmental students academically prepared to learn the contents of your course?
2. Have we taught appropriate skills and knowledge for developmental students to perform adequately in your class?
3. Are developmental students applying the skills and knowledge needed to learn in the courses in which they are enrolled?

Similarly, a survey for learners who completed the developmental program can be given to examine the long-range effectiveness of changes in affective behavior. The learner might be asked to respond to the following types of questions:

1. Do you feel more confident about your ability to read, write, study or solve problems? Why?
2. Are you interested in learning the subjects that you are studying this semester? How do you know?

3. Are you usually motivated to study for your courses?
4. When you are not motivated, what do you do to make yourself study?

Finally, all of us have probably experienced a developmental student who enthusiastically expresses the pleasure and delight of receiving an "A" on a test after receiving several "F"'s on prior tests. The student often attributes the "A" to the learning that occurred in the developmental program. For instance, "I never learned how to organize my notes to study, but now I do, thanks to the help you gave me." This type of evaluation is informal and sporadic, but it can be organized in such a way that meaningful information is collected if the practitioner records the "testimonial data" on a regular basis.

Summary

This section has provided a model for evaluating learning in developmental programs. The model demonstrates that evaluating learning occurs in three stages: evaluating the learner prior to instruction; evaluating the learner, the learning environment, and the learning during instruction; and evaluating the learning after instruction. Both affective and cognitive influences on learning are considered in the model as well. Finally, the model suggests many methods to evaluate learning, and most likely, practitioners will have to utilize a variety of them. The next three sections of this publication will present specific information on using the many tools that are available to do the job of evaluating learning.

SECTION II: TESTING IN DEVELOPMENTAL PROGRAMS

By Nancyanne Rabianski

Testing of the student's learning is an important part of any developmental program. Initially, a test or battery of tests, of a formal or informal nature, determines a student's placement in a developmental course(s) rather than the college-level course(s). Later, still more tests will probably diagnose the student's particular deficiencies, so that he/she will receive the proper instruction. During the semester, additional tests will monitor his/her achievement, and, finally, a test will determine the student's eventual exit from the developmental course.

Since the instructor (not to mention the student) spends a great deal of class time in testing and relies heavily on test results in making important educational decisions, the instructor needs to know as much as possible about not only the uses but also the weaknesses of various types of tests. This section will briefly explain the types of tests used to determine placement, provide diagnosis, and measure course-related achievement as well as discuss, in some detail, one specific example of each type of test. Additional frequently-used tests cited by Martha Maxwell (1980) and their availability will also be indicated. Perhaps even more importantly, the discussion will define some psychometric terms commonly used by admissions counselors and test publishers in their discussions of testing. A knowledge of such terms will assist the instructor in judging the effectiveness of his/her own tests and engaging in discussions with admissions counselors, administrators, and fellow instructors regarding testing within the developmental program.

Placement in the Developmental Program

Underprepared students who enter colleges and universities are not always placed in mandatory developmental programs, but those who are placed in such programs are usually sorted from the other college students on the basis of standardized tests, locally-built tests and/or informal instruments. Even though there is a variety of ways to evaluate and place students, STANDARDIZED TESTING is probably the one way used most widely. According to Martha Maxwell (1980), 95 percent of the college reading tests used to select, place and/or evaluate students are standardized. She also indicates that standardized tests are used in math, and a number of English programs, even though most English departments also require a writing sample as either an addition or an alternative to the standardized test.

Standardized tests are constructed by test publishers and designed for use in a wide variety of schools. To ensure broad coverage, the test makers consult widely used textbooks and the opinions of curriculum experts. The tests are standard in the sense that the directions for administering and scoring the test are very precise. Over a thousand standardized tests are in existence. Most of them are norm-referenced.

NORM-REFERENCED TESTS are useful in comparing students and sorting them into groups. A norm-referenced test is so called because it allows interpretation of the test score of a particular pupil in relation to other pupils of similar age and, perhaps, background. W. James Popham (1975) indicates that norm-referenced tests are indeed the best measure available when selections must be made.

In preparing norm-referenced tests, the publishers of the tests gather large groups of students of various age groups and from various geographical areas. The students take the tests under uniform conditions such as restricted time limits. After the students' responses have been gathered and tabulated, items that appear to discriminate among students are retained. Items that discriminate are those on which there are a range of student responses. Items are then refined and retested. Eventually the items are arranged so that there is a range of difficulty of items. The scores achieved by students, when arranged in a distribution, form a bell-shaped curve; that is, 50 percent of the students' scores fall above the average score (mean) and 50 percent fall below the average score. Test norms are commonly reported in the form of age norms, grade norms and/or percentile norms. Grade norms tell the grade group in which the pupil's raw score is average; age norms tell the age group in which the raw score is average; and the percentile norm tells the percentage of students in the reference group who fall below the pupil's raw score. Norms are also often reported on the basis of national, state and local groups, so that the person interpreting the student's raw score has a choice of the group with whom the student is compared.

Some colleges, for purposes of placement in developmental programs, do not use published norms. They simply establish cut-off scores on standardized tests. For example, at the University of New Haven, the Test of Standard Written English (TSWE), a subtest of the SAT, aids in the placement of composition students. The test requires the student to identify whether a section of a sentence is correct or needs to be replaced with a choice that contains a correction. The SAT's test items have been designed in the manner common to norm-referenced tests, and data regarding students who have taken the test are available, but norms have not been established. The publisher encourages test users to base their interpretations on the scores commonly found among their own students (Steve Ivens, 1981). Martha Maxwell (1980) also encourages the establishment of campus

norms since they are the best indicator of the range of student knowledge on the home campus. In the case of students at the University of New Haven, those students who receive a raw score that is near the cut-off point established by the English department must take an essay test in order to determine final placement in a developmental writing class. The initial placement is generally supported by the essay test score. The utility of the TSWE in placing students has been demonstrated in a 1975-1976 study; students from four major colleges took the TSWE before their first course (770 students), toward the end of the first course (776 students) and after a full year of instruction (457 students). Results indicated that the performance on the test paralleled the instructional level of placement (Breland, 1977).

In addition to providing normative data, standard tests also usually offer published reliability and validity figures in a manual that accompanies the test. Reliability involves the consistency of scores obtained when a student is examined with the same instrument or an equivalent measure on repeated occasions. However, high reliability does not always ensure that a test is worthwhile (Wallace and Larsen, 1978). Validity is more important in selecting a test because it involves the extent to which a test measures what it was designed to measure. Most standardized tests are accompanied by a manual that indicates validity and reliability data for the test.

According to Martha Maxwell, several standardized tests are commonly used as placement instruments:

Screening Tests:

Wide Range Achievement Test (WRAT). Jastak Assessment Systems, 1526 Gilpin Avenue, Wilmington, De. 19806. Reading (a one-minute test administered individually in which the student reads orally from a word list of increasing difficulty), spelling, arithmetic computation. Two levels: age 5-11; age 12-adult.

Carver R. Reading Progress Scale. Revrac Publications, 10 West Bridlespur Drive, Kansas City, Mo. 64114. A quick, eight-minute test for grades 3-16.

Longer Placement Batteries:

Adult Basic Education Student Survey. Follett Educational Corp., Chicago, Ill. Reading comprehension, word recognition, arithmetic computation, arithmetic problems.

Adult Basic Learning Examination. Harcourt, Brace, Jovanovich, Inc. Level 1 (ABLE 1), grades 1-4, reading, vocabulary, spelling, arithmetic problems; level 2 (ABLE 2), grades 5-8; level 3 (ABLE 3) for adults reading between grades 9 and 12.

Sequential Tests of Educational Progress (STEP). Cooperative Tests and Services, Educational Testing Service, Princeton, New Jersey. Reading, English expression, basic mathematics concepts,

science, social studies, grades 13-14.

McGraw Hill Basic Skills System. California Test Bureau/McGraw-Hill, Monterey, Ca. Separate tests in reading, study skills (including use of library), and mathematics, grades 11 through college. Reading subscores include recreational-reading rate, study-reading rate, flexibility, retention, skimming and scanning, paragraph completion.

Descriptive Tests of Language Skills. The College Entrance Examination Board, Princeton, New Jersey. Reading (understanding main ideas, understanding direct statements, drawing inferences), sentence structure (using complete sentences, using coordination and subordination, placing modifiers), logical relations (categorizing ideas, using appropriate connectives, making analogies, recognizing principles of organization), usage (pronouns, modifiers, diction, idioms, verbs), vocabulary. [These are new tests and norms being developed (Maxwell, 1980)].

General:

Comparative Guidance and Placement Test. Educational Testing Service, Princeton, New Jersey. Self-scoring. (Maxwell, 1980).

In order for an instructor to become familiar with such tests, he/she can order a specimen set of test materials from the publisher. In addition to the data contained in the manual, an instructor can learn more about a test by consulting O.K. Buros' **Mental Measurements Yearbook** (1978, latest edition) and **Tests in Print II.** Journals such as **The Reading Teacher, Journal of Educational Research,** and **Journal of Educational Measurement** also review tests, but Buros' volumes are the most comprehensive and accessible source of data and reviews. Buros lists information regarding the test such as the various subtests, testing time, and test prices. The yearbook also gives numerous bibliographic references to the test and appropriate reviews in their entirety.

For example, Buros' **Eighth Mental Measurement Yearbook** offers not only basic information but also reviews regarding the well-known Comprehensive Tests of Basic Skills (CTBS). The battery is designed for grades kindergarten through 12.9. Subtests are available for grades 8.5-12.9 in the areas of reading, language, mathematics, reference skills, science and social studies. One reviewer of the CTBS, Warren G. Findley (1978), has noted that the science and social studies subtests are a concession to curriculum authorities who want to develop basic skills without losing sight of science and social studies as basic elements of the curriculum. Findley, however, views the tests as being ill-defined tests composed of reading comprehension, chart interpretation, mechanical comprehension, basic information, and uncertainly specialized vocabulary. Nevertheless, he cites the entire battery of tests as being generally sound. Findley especially favors the groups of students used while developing the test norms; the group

was slightly overrepresented by minority ethnic groups, as he felt might be the case with students for whom the test was designed. An additional strength of the test battery cited by J. Stanley Ahmann (1972), another reviewer, is that the instructions and artwork are clear and easy to interpret.

The Comparative Guidance and Placement Test is another placement test that according to Martha Maxwell (1980), is of particular note because it is a self-scoring test that students can take while they are waiting to see an academic advisor. The battery includes 1. Interest and background inventories, which yield an academic motivation score, and interest scores for mathematics, physical science, engineering technology, biology, health, home economics, secretarial work, business, social sciences, fine arts and music and 2. achievement/placement tests that include reading, sentences, mathematics, the ability to follow directions, perceptual speed and accuracy, and inductive reasoning. The test is designed to be used by entrants to postsecondary institutions, particularly two-year institutions. Norman Eagle, a reviewer notes that overall the test represents a distinct advance in programs that claim to aid students in making postsecondary educational or career decisions. Nevertheless, the reviewer also gives some cautions regarding the test. For example, ETS has established practically no relationship between the interest scales and course grades except for a small relationship in the area of mathematics. In order to use the tests wisely, an instructor should consider such warnings as those offered by the reviewers in Buros' yearbooks.

The reviews, written by testing experts, reveal the overall purpose of the test as well as the test's strengths and weaknesses. For example, in a review of STEP II, one of the placement devices cited by Maxwell, the first reviewer, John Rosenbach, an educational psychologist, indicates that the purpose of STEP II, according to its manual, is to evaluate an individual's ability to apply so-called central learnings rather than to demonstrate knowledge or skills specific to certain courses. The battery is designed for grades 4-14. For all grades there are subtests in reading, English expression, mathematics, basic concepts, science and social studies; for grades 4-12, there are also additional tests in the mechanics of writing and mathematical computation. The subtests' times vary from 40 to 60 minutes. The handbook of the test gives information on the standardization and norming of the test.

In addition to such basic information, Rosenbach (1978) offers some warnings about the test. For example, he says that the college-level norm samples appear suspect because of the problem of obtaining the cooperation of institutions. He also warns that each test user should examine the actual test content with respect to his/her own educational objectives. He points out:

> The heavy overlap in content, particularly among social studies, reading, mathematics, basic concepts, and science tests is, in all likelihood, a reflection of the difficulties encountered when one attempts to construct an achievement test designed to measure higher order processes while avoiding more specific instructional content...such tests tend to take on the characteristics of general ability tests.

In other words, he feels that the test is too much like an aptitude test (test of ability) rather than an achievement test (test of knowledge). The second reviewer, Richard E. Schutz, (1978) Executive Director of SWRL Educational Research and Development, also complains that the test does not truly measure what the test manual refers to as "skills and understandings that should be part of the repertoire of every well-informed citizen." (He does, however, say that the STEP II series is a model of excellence if judged according to classical psychometric criteria, which are standards developed in relation to norm-referenced measures.) Both reviewers seem to be objecting to the very nature of the test as a norm-referenced instrument.

Norm-referenced instruments by their nature do have some limits to their usefulness. In developing such instruments, in order to ensure that there is a bell-shaped distribution of students (50 percent of the students above and 50 percent below the mean), some test makers eliminate questions that do not discriminate among students, even though the reason students may all respond correctly to the question is simply because the question reflects concepts stressed in the classroom. The items that discriminate among students and, therefore, tend to be retained are usually general and reflect aptitude (ability to learn) rather than achievement (knowledge already possessed). Consequently, the norms allow the comparison of students, but they do not yield an absolute score that tells how well a student has mastered a subject or skill. A criterion-referenced test, however, does just that. Criterion-referenced tests will be addressed briefly in the following discussion of diagnosis.

But what about normative-referenced testing? Many of the standardized tests have been developed and interpreted in this fashion. Should instructors simply throw away all the hundreds of norm-referenced tests? Probably not. The purpose of such instruments is to discriminate among students, and that is the purpose of placement. In regard to the concerns about the relevance of norm-referenced tests to the curriculum (Popham 1975), the instructor can determine the degree of relevance by asking the campus admissions officers for a specimen test and a manual. A specimen set will indicate whether or not the test items are compatible with the local curriculum. If the test reflects the curriculum, but an instructor continues to have doubts about the instrument because, for example, years of

observation have indicated a number of questionable placements of students, then the matter can be pursued further in Buros' reviews and ultimately, journal reviews. If, on the other hand, an instructor's observations have led him/her to believe that most placements are correct, then probably the standardized test, despite the weaknesses of norm-referencing, is performing its basic aim of sorting students and placing them in a relative standing.

Diagnosis of Deficiencies

Before instruction, a student's particular deficiencies must be diagnosed so that he/she receives the proper instruction. According to Thorndike and Hagen (1977), diagnostic tests, since their purpose is to determine specific areas of deficiency in an individual student, should be constructed in the manner of criterion-referenced tests. What is a CRITERION-REFERENCED TEST? In contrast to a norm-referenced test, which focuses on a student's relative standing within a group, a criterion-referenced instrument compares the student to a desired level of proficiency, a mastery level. For example, the General Educational Development (GED) tests use the actual achievement of high-school graduates to establish the minimum standard that must be reached in order for a student to obtain a GED high-school equivalency certificate. Although the development of criterion-referenced testing is relatively new (mid 1960's), interest in such tests has been increasing, and a number of standardized instruments of this sort are now available.

In designing a criterion-referenced test, the maker must first establish classes of expected learner behaviors, such as the ability to add correctly a pair of single-digit numbers. Items are developed which sample the class (domain) of behavior. Since the interpretation of the test score does not require a variance among the responses of students, items that reflect the objectives are retained regardless of the number of students who may answer the question correctly. Since the content of such tests reflects particular, stated behaviors rather than the more general, diffuse knowledge sampled by norm-referenced tests, they are useful in determining particular deficiencies of the learner.

Criterion-referenced tests, however, may have a deficiency of their own. According to Robert L. Ebel (1975), as one progresses to more advanced learnings, it becomes increasingly difficult to identify a limited number of specifies to be learned by mastery. There are also aspects of this type of testing that are not, perhaps, inherent weaknesses, but potential, related problems for the instructor. For example, the development and use of such instruments can be time consuming. The instructor must use parallel test forms to test students repeatedly if they do not reach mastery on the first test administration, and reports of the student's progress in each area may be voluminous.

The possibility also exists that teachers may teach too directly to the test if the goals listed are too restricted (Ebel, 1975 and Brown, 1976).

The Stanford Diagnostic Reading Test and the Stanford Diagnostic Mathematics Test are two standardized, diagnostic tests that Martha Maxwell suggests are commonly used in developmental classrooms. Both tests are available from Harcourt-Brace-Jovanovich, 757 Third Avenue, N.Y., N.Y. 10017. Both offer criterion-referenced interpretation of scores of students up to the level of grade 13. The reading test measures comprehension, vocabulary, decoding and rate. The mathematics test measures the number system and numeration, computation and applications.

Both tests have received some criticism from reviewers. Byron H. Van Roekel (1978), one reviewer of the reading test, states that it has "few peers among group diagnostic reading tests." However, he cautions that the objectives and sequence of tasks have been passed upon information gathered from reviewing curriculum materials, not from a theoretical model of reading acquisition, so a test user should carefully consider the test in relation to his/her own curriculum and students. A reviewer of the mathematics test, Glenda Lappan, also warns that the test user should check to ensure that the test's objectives and mastery levels reflect local expectations. In addition, Larry Sowder (1978), another reviewer of the mathematics test, raises a question that applies to the Stanford Diagnostic Mathematics Test, the Stanford Diagnostic Reading Test, and many other standardized tests as well that are supposedly criterion-referenced tests. Sowder points out that in addition to the mastery levels established for use in interpreting the scores, norms have also been established. Why would anyone want to use norms (scores for comparison purposes) in interpreting a diagnostic test? Sowder concludes that if an instructor wants somewhat detailed information on a narrow range of concepts or skills, a local test might be better than a standardized instrument. A detailed discussion of teacher-made criterion-referenced tests will appear in Section III of this publication.

Another important area of possible deficiency must be diagnosed in addition to the knowledge (achievement) deficiencies that are usually tested. In addition to knowing the area of deficiency, the instructor must know the reason for the deficiency before planning the instructional program. The reason is sometimes an underlying problem that can be eased with an understanding of the student's attitudes, habits or learning styles. Kurt Lauridsen, Director of the Student Learning Center at the University of California at Berkeley, in a longitudinal study recently concluded at Berkeley (1981) found that affective factors such as attitudes were very important in a student's success or lack of success. This recent study concurs with many cited by Martha Maxwell (1980). The studies cited characterize low

achievers as students who, for example, fail to assume responsibility for their own learning, do not receive emotional support from parents, lack personal goals and values, have poor or nonexistent study habits and are anxious. If these are indeed characteristics of students who do not achieve, then we need to test for such characteristics and help these students as much as possible.

Several measures are well known in the area of STUDY HABIT AND ATTITUDE measurement. One which is especially well known is the Brown-Holtzman **Survey of Study Habits and Attitudes,** which has been renamed **The Survey of Study Habits and Attitudes (SSHA).** SSHA is available from Psychological Corporation, N.Y., N.Y. The instrument is for use in grades 7-12 and 12-14. The inventory consists of 100 items. Four 25-item subscores are delineated: Work methods, Delay Avoidance, Teacher Approval, and Education Acceptance. Students are directed to respond to numbered statements, arranged in two columns per page by indicating one of five responses: rarely (0-15% of the time), sometimes (16%-35%), frequently (36%-65%), generally (66-85%), or almost always (86-100%). According to Albert E. Roark and Scott A. Harrington (1978), the SSHA may be a good diagnostic instrument, although a major weakness is that the scores can be manipulated by the student. The instrument relies on the student's frankness. Another frequently used survey of study habits, according to Martha Maxwell, is the SR/SE, a **Survey of Study Habits and Attitudes** for high school and college students, which is available from Science Research Associates, Inc., 259 East Erie Street, Chicago, IL 60611. When difficulties are diagnosed, Maxwell suggests supportive counseling/teaching and structured learning experiences. In addition to the diagnostic value of instruments such as the SSHA, they may be used as screening devices to identify students who need a study skills course.

COGNITIVE STYLE INVENTORIES also provide some insight into the needs of the student learner that must be met in order for him/her to be successful. Not all students learn best in the same way. Differing terminology is used in the psychological literature, but basically cognitive style, learning style or conceptual style all refer to a person's typical mode of perceiving, thinking, remembering and problem-solving. Martha Maxwell (1980) suggests that a simple assessment of cognitive style could identify appropriate study strategies for a student; for example, a student who is visually strong might make multi-colored charts of lists of facts that need to be remembered for a course.

One such well known instrument is the **Canfield Learning Styles Inventory**, which is available from Humanics Media, Liberty Drawer 7970, Ann Arbor, Michigan 48107. The inventory investigates the following areas: 1. concerns for the dynamics of the situation in which

learning occurs (including relations with others, clear study plans, independence of action, comparing self with other, and desire for authority) 2. major area of interest (including numeric, qualitative, inanimate and people-related choices) 3. general modality through which learning is preferred (including listening, reading, iconics and direct experience). Canfield's inventory is easy to administer and score. An alternative instrument is the I-S Scale (Impulsivity-Stability), found in D.E.P. Smith's **Learning to Learn** published by Harcourt-Brace-Jovanovich.

No matter how well a student's learning style is diagnosed and accommodated with appropriate studying techniques, and no matter how well a student has learned a particular subject matter, some students still fail because of test anxiety. Standardized tests are particularly likely to generate panic because unlike mastery learning, in which a number of parallel tests can be administered to a student, standardized testing is likely to create a "now or never" feeling in a student.

How can such test-anxious students be identified? Many, of course, are obvious from their self-reports in the form of loud complaints and from the teacher's observation of haggard, sleepless faces prior to an exam. A diagnostic instrument, however, is necessary in order to determine whether the anxiety is caused by a learned response to testing, a lack of study skills, a lack of knowledge regarding test taking skills, or poor classroom instruction. One test that Martha Maxwell (1980) suggests is R. Alpert and R.N. Haber's **Achievement Anxiety Questionnaire**, which yields two measures of test anxiety, facilitating anxiety and debilitating anxiety. The two measures are taken because the authors of the instrument recognize that some degree of anxiety is motivating, not debilitating. The questionnaire is available in a 1960 issue of **The Journal of Abnormal and Social Psychology** (61, pp. 207-215).

Achievement in Classes

In a broad sense, every paper and pencil test is an ACHIEVEMENT TEST because it measures the knowledge or skills that a student already possesses. Achievement tests can be broad, standardized tests of the sort discussed under placement, in which case, the score for the whole test may be used to compare students with each other. Achievement tests can also be used for diagnosis, if the subject scores are considered for an individual's mastery of material. This discussion, however, will focus on achievement tests that are teacher-made and designed to test achievement in the developmental course that a student is taking or from which he is attempting to exit. In other words, the domain under discussion will not be as broad as high school math. The domain (area of knowledge or skills) will be of the sort tested frequently during instruction (formative evaluation) and at the

conclusion of a semester of instruction (summative evaluation).

According to Frederick G. Brown (1976), there are several uses for achievement tests during the course of instruction. First, the most important use is to provide feedback to students. In order to learn, a student must know whether or not his/her responses were correct. Frequent testing and detailed feedback is time-consuming but necessary. Those of us who have used Hunter Boylan's form for reporting midterm progress (1979)[1] know that developmental students are eager to receive reports from their professors regarding their ongoing progress in a course. Secondly, tests can stimulate learning if they are used for the purpose of reinforcement. Third, achievement tests can be used for counseling purposes. If, for example, a standardized aptitude test (an indication of ability, such as IQ test) indicates that a student has potential that is not clear on the daily tests, the teacher can turn to the diagnostic tests of interests and study habits to try to discover the problem, and then try to speak with the student. Fourth, the test can provide the instructor with some feedback regarding the effectiveness of instruction. By studying the pattern of incorrect responses, the instructor can discover gaps and misconceptions on the part of the student's knowledge. According to Robert Fuller, in a recent issue of **The Journal of Developmental and Remedial Education**, "The wrong answer tells you more about the student and her/his reasoning than all the right answers in the world." (1981) Fifth, the tests can also be used as a check on students' studying. A quiz is one way of determining if students are understanding the required reading. Finally, tests can be used to stimulate discussion if the student's responses or difficulties with certain items are used as the basis for further explanation or discussions with the class (Brown, 1976).

An achievement test is also important at the conclusion of instruction (summative evaluation). It may be used to determine if the student has attained some kind of minimal proficiency necessary for exiting the course. The performance standards, of course, should be established early in the course. A parallel form of the test may be used as a pretest before instruction begins. A pretest can serve as a good, teacher-made diagnostic tool.

In order for achievement tests to be useful, they must be carefully devised. At least three major points should be kept in mind in designing TEACHER-MADE COMPETENCY TESTS. First, the content that is to be covered in the test must be specified in behavioral terms. Educational goals must be stated in behavioral terms (observable behaviors) in order for them to be measured. For example, a positive

[1]Boylan, H.R. "The Role of Academic Intervention." Paper presented at the 6th National Conference on Developmental Studies. Lexington, KY, April 4-6, 1979).

attitude toward learning can be inferred but not measured. Second, the test must measure the specified goals, not irrelevant considerations. For example, if the test is attempting to measure a student's ability to use algebraic formulas, then the test should not feature word problems because the student may not be able to solve the problem as a result of reading problems, even though he/she can use the algebraic formulas. The third point is that test takers must have had an opportunity to learn the material tested (Brown, 1976; Smith and Adams, 1966).

In the developmental classroom the design of the exit test, a test of basic competency, is especially crucial since evaluation of the test determines not only the student's satisfaction of the objectives of the developmental course but also his/her readiness for the college level courses that follow. The National Council of Teachers of English has responded to the need for good competency testing by publishing **The Nature and Measurement of Competency in English** (1981), a collection of essays that deal with defining and assessing competence in language, reading, writing and the media. For example, in the area of reading, Alan Purves (1981) points out that the testmaker must be careful to ensure that reading passages don't deal with subject matters, such as women's rights, of which a reader must already possess an understanding in order to understand the passage.

He also presents thought-provoking questions regarding the content of competency tests. Any credit buyer should be able to read and understand a Master Charge statement, or at least one might think so, until this everyday bit of reading is examined: "The balance on which the Finance Charge (if any) is computed is the Average Daily Balance, determined as follows: (i) the balance on the first day of any billing month..." An examination of a passage such as the preceding Master Charge statement makes one wonder just what is basic competency for a reader.

Ultimately, Purves suggests that a test in reading competency should include the understanding of words in context, the determination of meaning of a larger passage, the determination of the tone of a passage. A cloze test might measure the words in context. The meaning of a larger passage might be tested with several short quotations supplemented by one or two paragraphs of brief length. The determination of intention might be measured with the matching of a picture to a description, and the measure of tone might be in a short answer format. Similar, specific points regarding the nature of competency and its assessment are made in the essays dealing with language, writing and media. The book of essays is available from the National Council of Teachers of English, 1111 Kenyon Road, Urbana, IL 61801.

Helpful information regarding the design of tests in the area of mathematics is available from the Mathematical Association of

America (1529 Eighteenth Street, N.W., Washington, D.C. 20036). The MAA publishes a newsletter that discusses testing issues.

Summary

If the tests used on a particular campus seem to be serving a useful purpose in placing students or diagnosing difficulties of students in the program, they may be serving their purpose. If, however, an instructor feels suspicious of the tests, he/she should take the time to check the content to ensure that the goals of the test indicated in the manual match the purposes for which it's being used. Also, the instructor should check to see that the objectives being tested by the items match the objectives of the local curriculum. The instuctor should not be complacent about an issue as important as testing. Thousands of tests are available, and Buros' **Mental Measurements** and **Tests in Print** are good guides to choosing amongst them. Above all, the instructor should not neglect his/her own test-making potential to develop clear objectives and mastery levels for his/her students. They deserve to be tested often and given clear, detailed feedback as they are expected to learn.

SECTION III: DEVELOPING AND USING CRITERION-REFERENCED MEASUREMENTS IN DEVELOPMENTAL PROGRAMS

By Karen G. Smith

Criterion-referenced testing may become recognized as the most sensible and meaningful change in procedures adopted by educators in the 80's. Actually, the term "criterion-referenced test" was created and defined in 1963 by Glaser as a test deliberately constructed to yield measurements directly interpretable in terms of specified performance standards. Criterion-referenced measures are those which are used to ascertain an individual's status with respect to some criterion or performance standard (Cartier 1968; Glaser 1963; Greco 1974; Popham and Husek 1967; Smith 1973).

Since Glaser's conceptualization of criterion-referenced testing, many educational researchers, theorists, and practioners have developed, used, refined, evaluated, and critiqued its merits and faults. Those of us in developmental programs have been among the most fervent proponents of this method of learner evaluation.

Popham (1978) defines "criterion" in criterion-referenced testing as an instructional objective, a desired level of learner proficiency, or a desired post-instruction outcome. Criterion-referenced tests are used to identify an individual's status with respect to an established standard of performance. Performance on test items is meant to tell how much the student knows about the topic or how well he or she can perform the task (Roberson 1975).

Appropriateness to Developmental Education

Criterion-referenced measurement is a direct outgrowth of the need to adequately assess learner performance on programmed instruction. Standardized tests which were available in the late 50's and early 60's, when programmed instruction developed, did not work well in measuring the results of instructional sequences which were designed to lead the majority of learners to mastery of the instructional objectives. Standardized tests were too general, and it was obvious that instruments designed to sort out learners across a normal distribution curve could not adequately serve in instructional programs designed to achieve common mastery.

Therefore, the movement to individualized instruction fostered the comparable trend toward criterion-referenced assessment, since the introduction of mastery learning and individualized instruction forced educators to identify more appropriate assessment devices. Especially

true in developmental education, the use of traditionally developed norm-referenced tests yields inappropriate and inadequate information for both diagnosis and the assessment of learning outcomes.

Our task in developmental programs is to determine the skill levels of our students, the appropriate point of departure for the instructional sequence, and the student attainment of the instructional objectives. Standardized tests cannot adequately diagnose or identify specific skill attainments, but, instead, yield information generated from the norming procedures with traditional populations — percentile rankings, stanine groupings, or estimated grade level equivalents. Norm-referenced tests refer to or get their meaning from a ranking or a norm. They do not tell us what or how much a student knows. They do tell us where he ranks with reference to other students.

Commercially prepared standardized tests can, however, provide some degree of information about the individual student's progress if a systematic item analysis is applied to the test. The purpose of an item analysis is to see which test items match the learning outcomes previously identified and what the proportion of those relevant test items is to the test as a whole. Typically, the item analysis will demonstrate that too few, if any, of the items in the standardized test describe performance behaviors that match the instructional objectives and outcomes desired. Too often, there is a substantial lack of congruence between what the norm-referenced test measures and what is stressed in the curriculum.

Gronlund (1973, p. 8) states, "Criterion-referenced tests are easiest to design, construct and use at the mastery level of learning." This is true because the learning outcomes are relatively simple, the domain of behavior is limited, the learning is sequential, and the percentage-correct score indicates degrees of progress toward mastery. As stated in the previous section, beyond the mastery level, however, some difficulties arise with the use of criterion-referenced tests. The learning outcomes are more complex, the domain of learning tasks becomes unlimited, and the learning seldom proceeds through a neat sequence of stages. Instructional goals at the developmental level often represent goals to work toward rather than goals to be fully achieved, for we are striving at this level for continuous development of understanding and skills.

Shortcomings, notwithstanding, the well-developed criterion-referenced test provides the best and most usable information for the post-secondary remedial-developmental educator.

Development of Criterion-Referenced Tests

The development of criterion-referenced tests should lead toward measures of achievement that express student performances on

clearly defined educational tasks. Although more comprehensive guidance can be found in the literature, the following principles, derived generally from Gronlund (1973), Popham (1978), and Roberson (1975), can provide a general framework for criterion-referenced test development.

1. **Identification and delimiting the instruction area to be tested.** One way to limit the area to be tested is to divide the instructional domain, or course, into relatively small units. These units should be coherent in topic, theme, skill, or some other segment of the instruction. Since the test to be developed will be very specific to the instructional domain, achievement results wll be reported in terms of what the student has or has not mastered in that domain of the instruction.

In any particular achievement area, it is easier to define the instructional domain for the lower level outcomes such as memory of facts than for the more complex ones of synthesis, application, and evaluation.

2. **Identification of the learning outcomes.** It is common practice to develop instructional objectives in order to systematically organize the instructional program and monitor progress toward those objectives. Tuinman (1978) states that a priority in criterion-referenced measurement is to establish first those objectives which contribute to or are prerequisites for the global goals to be achieved. Then the instructional objectives must be defined in terms of learning outcomes. In other words, each instructional objective should be more specifically defined by listing the specific learning tasks on which students will demonstrate performance at the end of the instructional unit to show that they achieved the objective.

Gronlund (1973, p. 18) provides the following example of an objective and related learning outcomes:

Understands scientific principles

1. States the principle in his own words.
2. Identifies an example of the principle.
3. Identifies the relationship between two principles.
4. Predicts an outcome based on the principle.
5. Formulates hypotheses in harmony with the principle.
6. Distinguishes between correct and incorrect applications of the principle.

The list of specific learning outcomes could be more exhaustive, but the instructor-evaluator's goal is to identify a reasonable number of learning outcomes which will provide a clear description of the objective. We could assume that how a student performs on these particular tasks would indicate how well he would perform on other tasks describing the same objective. More outcomes could be added, but a reasonable and representative sample is adequate.

The specific learning outcomes, or learning tasks, listed beneath

each instructional objective should be stated with an action verb that identifies the observable behavior. These performance terms describe the precise type of behavior the student will demonstrate as evidence of his achievement of the objective. The following action verbs give an example of those which describe the learning task; adds (multiplies, etc.), identifies, lists, matches, writes, states, distinguishes between, solves, shows, demonstrates.

The instructional objectives and identified learning outcomes for each objective comprise the test specifications. The test specifications, or rules to generate test items, are essential for the test developer. The learning outcomes statements, written in action format, provide a clear guide to develop test items which will be directly related to the outcomes, and thus describes the objective.

In basic skills areas and for diagnostic purposes, the objectives must be analyzed further and the learning outcomes defined to a finer point. For instance, "adds whole numbers," might be defined into even more specific tasks, as identified by Gronlund (1975, p. 11):

1. Adds two single-digit numbers with sums of ten or less (2 + 5).
2. Adds two single-digit numbers with sums greater than ten (6 + 8).
3. Adds three single-digit numbers with sums of ten or less (2 + 4 + 3).
4. Adds three single-digit numbers with sums greater than ten (7 + 5 + 9).
5. Adds two two-digit numbers without carrying (21 + 34).
6. Adds two two-digit numbers with simple carrying (36 + 27).
7. Adds two two-digit numbers with carrying into 9 (57 + 48).
8. Adds two or more three-digit numbers with repeated carrying (687 + 839).

Gronlund (1973, p. 18) provides us with three questions to serve as guides in determining the appropriateness of the learning outcomes to the objective:

1. Is each specific outcome stated in behavioral terms?
2. Is each specific outcome relevant to the objective?
3. Do the specific outcomes provide a representative sample of the behavior encompassed by the objective?

3. **Setting standards of performance.** Little specific guidance is available from the research to aid the criterion-referenced test developer in setting standards of performance. Recommendations have been made by many, but Cartier (1968), Gilman (1974), and Popham (1978) agree that the level of standard must be designated by the test constructor with full awareness of the ability of his students and the importance of the behavior they are required to demonstrate.

Haladyna (1976) advocates setting a confidence interval around a passing standard and making decisions based on how each student scores with respect to that confidence interval. Hambleton (1978)

claims that the most important question is whether or not the individual student attains some predescribed degree of competence on an instructional performance task. This is referred to as "critical criteria" by Tuinman (1978), who says the standard should clearly separate those students who will probably succeed on the next task from those who are likely not to succeed.

Tuinman (1978) advocates the use of two score results. The first, a mastery score, is obtained from a subset of terms for which all students have sufficient time to finish. The second score, the efficiency score, is in effect a speed score. He feels that efficiency is of particular interest in those tests which involve application and integration of lower order skills.

Block (1971) has suggested that 80 to 85 percent mastery is a realistic standard. He noted that setting the standard too high may be wasteful of teacher and student time and may have a negative effect on student motivation. The lower, but still high, standard tends to provide more opportunties for student success and for positive reinforcement.

Since the empirical evidence is not available to support any given level of mastery in classroom instruction, it is best to arbitrarily set a standard and then adjust it up or down as experience and circumstances dictate. The important decision in setting the standard of performance is based on the level of mastery necessary in order to learn effectively at the next stage of instruction.

4. **Generation of test items.** A test is a sample of tasks that was selected to represent some larger domain of behavior. If we are to generalize from the specific tasks in the test to the larger domain of behavior, then we must have some assurance that the sample of tasks is truly representative. Representative sampling is best assured when the area to be tested has been delimited to a specific domain (or concise unit of instruction) and when the learning outcomes to be measured have been well defined.

Gronlund (1973) suggests that the number of test items needed to get a representative sampling can best be determined by the use of a table of specifications. The domain's learning outcomes are listed across the top and the areas of content down the side. Each cell in the table can then be used to specify the number of test items needed for adequate sampling of the larger domain of behavior. "As a rule of thumb, it would be desirable to have at least several test items for each specific type of performance to be described and ten or more items for each general instructional objective," cites Gronlund (1973, p. 14).

It is important to note, at this point, that the difficulty of test items should be derived directly from the learning task to be measured. If none or few of the students can answer an item before instruction but all or most of them can answer it after instruction, both the test item and the instruction have been effective. Neither discrimination between

high and low achievers nor a spread of scores is the intent of mastery learning or of the criterion-referenced test.

In norm-referenced tests, the items are deliberately designed to rank students in order of achievement. One of the characertistics of norm-referenced tests is that in order to permit comparisons between an individual's performance and that of a norm group, the test must yield considerable response variance. Items which do not differentiate between high and low achievers are eliminated.

On the other hand, any item serves a useful purpose in the criterion-referenced assessment as long as it assesses an important attribute of the criterion. The difficulty of the items should be determined solely by the depth of the concept being tested.

Item form decisions should be tied to the learning outcomes statements. Since each test item should be designed to measure the behavior described in a particular learning outcome as directly as possible, the action verb used will most often dictate the item type needed in the test. Gronlund (1973) separates all item types into two major types; supply types and selection types. Short-answer and essay are supply types and multiple-choice, true-false, and matching are selection types of item forms. Any item form **may be** appropriate in the criterion-referenced test, but the learning outcome, if stated precisely, will specify which item form **is** appropriate.

5. **Concerns about validity.** Many critics of criterion-referenced measurement find fault with the fact that the traditional statistical procedures for determining validity and reliability in norm-referenced tests cannot be applied, and, therefore, the tests have limited worth and are not valid educational assessment tools. Other educational researchers and theorists, such as W. J. Popham, are proponents of criterion-referenced measurement and have been recently developing and reporting reliability and validity procedures. The researchers are not in general agreement, at this date, on statistical procedures for determining reliability. There is even some disagreement about the determination of test validity.

For our purposes, however, Popham (1978) provides guidelines for criterion-referenced test validity determinations. Functional validity, he states, is the accuracy with which the test satisfies the purpose to which it is being put. Functional validity is almost identical to criterion-referenced validity, and Popham claims that it is usually sufficient to know that the test possesses functional, or descriptive, validity.

As criterion-referenced assessment becomes more widely recognized, used, and researched, the theory and evaluation of instruments will become more sophisticated.

Uses of Criterion-Referenced Tests

Initially, tests were developed as instruments of assessment. They

were tools used almost solely to determine who had learned the material, best, least and in-between. Norm-referenced tests were developed during World War I by and for the military who needed quick indicators of achievement in order to determine which men would be named as officers. The American public school systems soon adopted norm-referenced tests, and their use has been almost exclusively for sorting out learners and for identifying specific groups of learners.

Criterion-referenced tests, on the other hand, are well suited to several educational purposes. Since they are discriptors of behavior in a relatively small or concise domain, they serve well as pretests or diagnostic instruments, progress indicators, and final assessment tools.

1. **Pretesting.** Pretesting will determine what prerequisite skills the student possesses and where the student should be placed in an instructional sequence. Pretesting will determine what portion of an instructional unit the student has already mastered and provides a base for measuring learning gain during the instruction.

The criterion-reference test is an excellent instrument for providing a description of the knowledge and skills possessed by the students which enables the instructor to provide any needed remedial instruction, to place students in the proper learning materials and learning groups, and to modify the instruction to fit the students' strengths and weaknesses.

2. **Formative testing.** Testing done during the instructional sequence is formative testing. Criterion-referenced tests are well-suited to this purpose. Formative testing assumes, however, that the tests are to be used to improve student learning and instruction and not to assign grades. Since each student can be recycled through the material until he has mastered it, the assignment of grades to the formative test results becomes meaningless.

The following steps, modified from Bloom (1971), provide a usable procedure for using criterion-referenced tests as an integral part of the instructional sequence:

1. Administer a criterion-referenced test at the end of each unit of instruction.
2. Analyze the results to determine which objectives each student has mastered.
3. Where an objective has not been mastered, examine the individual test items to identify the student's particular learning deficiencies.
4. Prescribe for each student the particular instructional materials and procedures he might use to correct his learning deficiencies.
5. Retest each student with another form of the test after he has had sufficient time to correct his learning deficiencies.
6. Use the information from the criterion-referenced testing to

improve instruction.

3. **Summative testing.** Tests given at the end of an instructional unit or at the end of the course and used primarily for assigning student grades represent summative testing. Aggregate information derived from the summative test may also provide the instructor with additional indicators of degrees of success or failure of the instructional sequence, methodology, and materials.

All items should be arranged by objective on the summative test. This arrangement makes it possible to score and interpret each part of the test separately, by objective. Interpretation of the test parts can provide the best information for evaluating the effectiveness or ineffectiveness of the specific instructional objective, the related learning outcomes, and the instructional sequence itself.

Summary

Some concerns and qualifications about the use of criterion-referenced tests need to be reiterated.

If the developmental instuctor plans to use criterion-referenced tests, whether as assessment of mastery only or for pretesting, formative testing, and summative testing, the procedures of delimiting the instructional domain and defining the instructional objectives and learning outcomes are crucial. Additionally, we are reminded by Tuinman (1978) that we must establish, within the confines of a particular instructional system, which objectives contribute to or are prerequisite for the global goals to be achieved.

Teachers have used elements of criterion-referenced testing for decades, but the use of such tests to measure carefully selected samples of clearly defined learning outcomes is relatively new. Since theorists and researchers have only just begun to develop information to guide the practitioner, we must face the problems of adequately defining the domain of behavior, of constructing a representative sample of learning outcomes, and of construction relevant test items in an approximate manner. Also, until a more adequate basis is provided for determining standards, the practitioner must use arbitrary judgment in setting standards of performance. We will make tentative judgments concerning instructional objectives, learning outcomes, test items, and performance standards, and we will revise and refine these judgments as experience and new information become available.

In spite of the shortcomings of criterion-referenced measurement, educators must be concerned with continuing the development of good criterion-referenced tests. If we are to do a better job of teaching basic skills to developmental learners, we must inprove our abilities to match methods and materials to our students. In order to do this, we need instruments which are very different from the usual standardized

achievement test, and we need instruments which are carefully planned, developed and tested. Good criterion-referenced tests can give us the information we need to define the learning needs of individual learners and to plan appropriate instruction.

SECTION IV: USING NATURALISTIC METHODOLOGIES TO EVALUATE LEARNING IN DEVELOPMENTAL PROGRAMS

by Carol Clymer Walvekar

Sections II and III of this booklet elaborate on ways in which learning can be measured by accumulating data in a fairly consistent and systematic fashion. In these sections, the authors present information on acquiring data with the use of some type of measurement instrument which yields some type of quantitative information. Specifically, placement, diagnostic, achievement, and criterion-referenced tests provide the evaluator with numerical data. This type of data is often necessary when a program administrator is forced to deal with large enrollments in developmental classes, inquisitive administrators who want a quick report of the program's apparent effectiveness, or scrutinizing federal agents who want to know how effectively the government is spending its grant or title money.

Typically, norm-referenced tests are administered in a pretest-postest format. Thus, the evaluator tries to assess what the learner knows prior to instruction and after instruction. Diagnostic tests and teacher-made competency tests, on the other hand, are frequently given as needed, essentially whenever the instructor needs more information about the learner and the learning. Criterion-referenced tests can be administered as an integral part of the instructional sequence. Usually, all tests are given because the instructor, for some reason or another, needs to quantify information about the learning. However, there are many times between the pretest and the postest or the diagnostic and competency test when the developmental staff must use observations, conferences, and records to assess learning. These tools are called naturalistic methods of evaluation, and in the next several pages, details of the value and utilization of these methods is presented.

Naturalistic Inquiry: History and Definition

The term naturalistic inquiry may be new to some practitioners in developmental programs, but it is a term that is well-known to researchers in many of the social science disciplines. Patton (1980, p. 41) reveals that the practice of understanding "naturally occuring phenomena in their naturally occuring states" was derived from the disciplines of anthropology, sociology, and psychology. These disciplines, in an attempt to understand the meaning of human behavior, believed that they needed research methods which were different from the agricultural or natural sciences "because people are different from plants and animals" (p. 44). The concept of **verstehen** developed as a result of this need. Patton (p. 45) defines **verstehen** as a

tradition which "stresses understanding that focuses on the **meaning** of human behavior, the context of social interation, an **empathetic** understanding based on subjective states and behavior. The tradition of **verstehen** or understanding places emphasis on the human capacity to know and understand others through sympathetic introspection and reflection from detailed description and observation." **Verstehen** is the essence of naturalistic inquiry.

Description and observation are two key evaluation methodologies used in naturalistic inquiry. These methodologies yield qualitative data, which is central to naturalistic inquiry. Patton (p. 22) describes qualitative data as "raw data from the empirical world" consisting of "...**detailed descriptions** of situations, events, people, interactions, and observed behaviors; **direct quotations** from people about their experiences, attitudes, beliefs, and thoughts; and excerpts or entire passages from documents, correspondence, records, and case histories."

Also central to naturalistic inquiry is that the evaluation, design, and variables are not predetermined. Data are collected in an open-ended fashion. The evaluator does not manipulate the research setting, but looks for patterns to emerge and determines every phase of the evaluation design based upon the preceding phase. Of course, evaluation questions and purposes are decided beforehand and issues and concerns are identified prior to collecting data, but the evaluation design does not have a hypothesis which requires specification of variables so that the researcher can "discover" the "truth". Guba and Lincoln (1981) note that the meeting tests of rigor, that is, reliability, validity, generalizability, and objectivity are not considered by the naturalistic inquirer. Consequently, doubts concerning the truth value, applicability, consistency, and neutrality of naturalistic research have been raised by scientific researchers. But Guba and Lincoln (1981), in their book **Effective Evaluation** discuss a variety of solutions that can be used to overcome these evaluation problems.

In sum, naturalistic inquiry is dependent on the collection of qualitative data; that is, data from the **real** and **natural** work of people; data that reveal human perceptions, feelings, and interactions; and data that is collected with as few constraints as possible on the individuals who are providing the data. Qualitative data is obtained through qualitative evaluation methods such as observing and interviewing. However, Guba and Lincoln (1981, p. 64) warn that it is "a gross error to equate quantitative methods with the scientific paradigm and qualitative methods with the naturalistic paradigm. Yet, there is a strong relationship in that direction in practice, that is, naturalistic inquirers lean most heavily on qualitative methods, while scientific inquirers lean most heavily on quantitative methods."

Using Naturalistic Inquiry

There is an evaluation technique used by composition instructors to determine patterns of spelling errors that students make. It is a very simple and straightforward technique. The student is asked to keep a spelling log of errors. In one column of the log, the student must record every spelling error from every paper that he or she has written. The student enters, in the next column, the correct spelling of each word. In the third column, the student is required to write down the reason for the misspelling. The student derives the reason by analyzing both spellings and determining which part or parts of the word are problematic. For example, many students misspell **grammar** because it is pronounced **grammer**. Hence, the student would write, in the third column, "the word is pronounced differently than it is spelled." In the final column, the student is asked to write down some method for remembering the correctly spelled word. Using the example above, the student might develop a mnemonic device such as "I learned grammar at gramma's."

I learned of the spelling log from a colleague who had a senior-level college student in his remedial English composition class. This student happened to be an outstanding art major and had managed to avoid freshman composition for all the years he was enrolled in the university. By the time he was a senior, he could no longer avoid the class. (Now students are not allowed to enroll in other courses until they pass freshman composition, but at that time there was no such policy.) He had an apparently severe spelling problem, so severe that his writing was totally illegible, an interesting characteristic for an outstanding artist. The severity of his spelling problem had caused several low grades on essay exams he had taken during his junior year. He visited the learning center for assistance. I looked at his papers and immediately administered a diagnostic test. We worked together on his spelling for about one year and made some progress, but very little. As a senior, he had to complete freshman comp, or no art degree. He was intent and enrolled in the lowest level composition class. In a few short weeks, the instructor of the class identified, with the use of the spelling log, three serious deficiencies that had caused the majority of his spelling problems. By the end of the semester, the student eliminated most of the spelling errors from his papers, was ready for the next level of composition and has since graduated.

The purpose of this lengthy example is twofold: First, no test, no matter how comprehensive, reliable, or valid, is sensitive to all learning difficulties. Therefore, we must incorporate some other, more sensitive, evaluation methodologies to evaluate learning. Second, the spelling log is an example of a naturalistic evaluation method. It is used in the real and natural world of the learner, it identifies the real and natural mistakes of the learner and it provides real and natural data that

reveal patterns of difficulty.

The spelling log is one small example of a naturalistic or qualitative evaluation method that can be used in evaluating learning in developmental programs. There are many more possiblities of the application of this area to evaluation in developmental education. Before discussing these possibilities, it is necessary to outline some basic principles involved in naturalistic inquiry. **The Process of Naturalistic Inquiry.** Guba and Lincoln (1981, p.92) report that naturalistic evaluation is basically a two stage process:

> "identification of concerns and issues of the audiences involved, and collection of information bearing upon those identified concerns and issues."

Risking oversimplification, they define "concerns" as matters of interest or importance that need attention or substantiation, and "issues" as areas of disagreement. Determining concerns and issues is basically three separate processes according to Guba and Lincoln.

The first process is called categorizing. It is not an easy step because it takes time to develop an ample categorizing system in that the concerns and issues can change before all categories can be established. At any rate, the categories should be based on observations, documents and records that the evaluator carefully analyzes to identify recurring regularities. The regularities help the evaluator to sort the information into concerns and issues which eventually become the titles of the categories. Sometimes it is difficult to put items into a particular category, hence items are labelled "miscellaneous." Guba and Lincoln feel that this is not a problem unless too many items fall into the miscellaneous category. If this happens, there is probably some fault with the categorizing system. Nevertheless, it is possible for an item to occur only once, seeming remote from the issue or concern, but still be important. Hence, subjective judgements have to be made in the categorizing process. Below are some guidelines that Guba and Lincoln (pp. 93-4) suggest for determining categories.

1. Check for internal homogenity among items within categories and external heterogenity among categories.
2. Categories may not be comprehensive or sequential since the data for categories is constantly arriving. The categorizing system is thus dynamic and changing any may be difficult to manage, but this is a characteristic of the real world and a part of naturalistic inquiry.
3. Categories should not be too broad or narrow. If the category is too broad, significant concerns and issues may be buried; if too narrow, the focus will be on detail and not on the total picture of the evaluation.

Once categories have been determined, they must be prioritized

because not every category can or should be evaluated. Guba and Lincoln (p. 94-96) identify seven ways to select priorities:

1. frequency of identified concerens or issues,
2. uniqueness of concerns and issues,
3. credibility in the concerns and issues by the audience,
4. ability of the concerns or issues to open up a new area of inquiry,
5. feasibility of evaluating the concern or issue,
6. special interest in a particular concern or issue,
7. materiality of the concern or issue (importance to the evaluand).

Finally, the evaluator must determine the completeness of the categories. In order to do this, Guba and Lincoln (p. 96-97) recommend that the set of categories should be:

1. internally and externally plausible,
2. of reasonably inclusive data,
3. reproducible by a competent judge (that is, verified for their existence),
4. credible to people who provided the data.

The categories that are developed will serve as a guide for Stage II of the naturalistic evaluation process: collecting detailed information regarding the concerns and issues. There are three steps outlined by Guba and Lincoln (p. 97) for completing the second stage:

1. using strategies to move from what is known to what is unknown,
2. employing criteria to make decisions about which to include or exclude,
3. deciding when to stop collecting information.

There are three strategies for determining what is unknown. First, the evaluator develops questions in order to access information that will extend what is already known. Second, the evaluator tries to bridge the information that has been collected within categories to make connections and discover interactions. Third, based on the known information, the evaluator proposes what new information should be found then looks for it. Guba and Lincoln (p. 99-100) state the criteria for deciding which information to include:

1. Is it pertinent?
2. Does it connect several existing information items?
3. Does it identify new pertinent elements?
4. Does it reinforce, but not replicate, existing information?
5. Does it explain other known information?
6. Does it exemplify the nature of the category or information in the category?
7. Does it refute or challenge existing information?

Finally, Guba and Lincoln (p. 100-101) say that one stops collecting information when there is nothing new, it is redundant, the area feels integrated and regularities are sensed, and finally, the information

seems remote from the concern.

Procedures and Methods for Collecting Qualitative Data. Up to this point, I have been trying to convey the basic structure of naturalistic research. It may seem a bit complicated, but then dealing with the real world and trying to evaluate the human beings that occupy it is a complicated task. Unfortunately, discussing procedures for conducting a naturalistic evaluation is only a little less complicated. Remember earlier I mentioned that the evaluation design of naturalistic inquiry is not predetermined. It is true that concerns and issues are decided before categories are established, but strategies for collecting information are not. This does not mean that naturalistic inquiry is a thoughtless endeavor. The evaluator must have some purposes, objectives, questions, and boundaries in mind or the evaluation will be mindless. However, specification of procedures and methodologies before the evaluation takes place is difficult because "each phase of the evaluation should build upon the preceding steps" (Guba and Lincoln, p. 275). Consider the following example: A program administrator may observe that an unusually large number of developmental course "add and drop" slips are being processed. This observation may signal that there is a "concern" with the placement tests. Consequently, the program administrator would want to find out if the placement test is accurately placing students in the appropriate developmental course. The administrator decides to interview students and faculty. Data from student interviews reveal that students have extremely negative feelings about the testing conditions. For example, the person who administers the test is often late, so students are not alloted the proper amount of time to take the tests. Thus, the concern of the evaluation might change somewhat and the administrator might begin to make observations of the actual testing situation.

Another important element of naturalistic inquiry is that the evaluator must be adept in using the methods of: observation, interviewing, and recording and analyzing data. Further beneath each umbrella of these three methods a variety of techniques must be effectively employed. Depending on the situation, an interview can be unstructured and spontaneous, it can be moderately structured with some outline of what is to be covered, or it can be standardized with specific questions that are answered in an open-ended fashion. Any of these approaches can be used. The vital issue is deciding which method will elicit the maximum amount of information with the least amount of disruption to the natural flow of things. For example, writing down notes constantly while conducting an interview can be very disturbing to the interviewee and thereby stifle a response.

The reader is dircted to the Patton (1980) and Guba and Lincoln (1981) texts for more details on the variety of naturalistic methods that

are available to the evaluator and how to use them.

Examples of Naturalistic Evaluation Used in Evaluating Learning.

Observation. 1. A program director wants to implement an ongoing method for collecting quantitative data about the learning that occurs in the developmental program. An outside consultant trains personnel to observe and record everyday occurences that might reveal information about the program concerns and issues. A tape recorder is kept in a central location for staff to quickly report their observations so they don't have to write lengthy reports. The staff is then asked to summarize interesting observations at weekly staff meetings. These summaries are the basis for issues and concerns that are evaluated later. The tape recorder provides more extensive documentation for determining items that will be used for categorizing and prioritizing issues and concerns.

2. Program personnel do not agree that the developmental program is meeting the needs of students. An evaluator acts as a student in the developmental program for two days and tries to covertly observe and participate in the program. The evaluator listens to the comments made by students and tries to find out their feelings about themselves and what they are learning. If no revealing, spontaneous comments are made, the evaluator engages students in a conversation about their perceptions. The evaluator looks for non-verbal cues and listens for verbal clues that reveal student's emotions and tries to find out the causes for these emotions. The evaluator observes interactions between students and teachers and student's interactions with the learning materials and other students. The evaluator records everything that is observed and heard and analyzes the observations according to the categories of issues that are being evaluated.

3. The developmental faculty is concerned about the quality of the testing program. An observer from outside of the developmental program participates in the placement testing program. The observer takes the tests that students are required to take and makes notes of observations. The observer is looking for answers to a specific set of questions. Is the testing environment conducive to taking tests? Are test directions given clearly? Is the noise level appropriate? Is the room comfortable? Is the time the tests are given scheduled during a good time of the day and is there ample time to take the tests? Are the tests appropriate? Observations are recorded and analyzed according to the categories of the concerns.

Interviews. 1. The developmental staff want to know if the program is meeting the affective needs of the learner. Each developmental faculty member holds individual conferences with every student

at least twice a semester. The same open-ended questionnaire is used by the staff to interview students about their attitudes, interests, motives, feelings, and difficulties that relate to learning. The staff records the students' responses at the beginning and end of each semester. Answers are compared and analyzed to determine differences, changes, and patterns. The responses are kept to signal or document issues or concerns that might be evaluated at some other time.

2. The director of the program wants to find out if the faculty and administration outside of the program feel that the developmental students are prepared to learning in non-developmental programs. The director discusses this issue with as many individuals as possible, records their responses and generates data which indicate that faculty and administration are concerned about this issue. The director decides to see if these findings can be documented with institutional records.
3. Program personnel want to find out about student's perceptions and history of learning skills at the time they enter a program prior to giving tests or looking at past records. An initial interview form is developed for all beginning developmental students. The interview is conducted by trained clerical staff. Some items on the form require specific responses and others are open-ended. The data collected on this form are used to make decisions about which placement tests the students will be given or in which parts of the program they will enroll.

Record Keeping and Data Collection. 1. In a developmental learning lab, the staff want to maintain a case study on the learning activities of each student who uses the lab. A non-confidential folder is kept for each student, who has access to it whenever it is needed. The folder is used by both the instructor and the student to keep all testing and instructional records and an information sheet. On the information sheet, the student writes down questions, concerns, difficulties and accomplishments while working in the lab. Every week, the instructor reads these comments and responds to them with other comments, answers to questions, and other assignments. This sheet provides a natural interactive evaluation system between the student and the teachers. It also produces a case history that can be used as a source for future naturalistic inquiries.

2. A program director wants to find out the rate at which developmental students are graduating from the institution compared to non-developmental students. Every semester, the director keeps very detailed records of students' names, addresses, i.d. numbers, dates of enrollment, and admission status. This data is compared to similar data that is obtained for

non-developmental students from the registrar's office. Each semester the director analyzes the data to see if patterns or regularities emerge, and waits to see if the data will signal a need for other evaluation studies.

3. Developmental faculty want students to have the opportunity to evaluate the program anonymously and spontaneously. They also want a written record of the students' evaluations. A suggestion and comment box is kept at a central location and students are encouraged to use it. Once a week, the comments are taken out and placed into categories which are considered for future action.

Summary

Of course, naturalistic evaluation has limitations. Data collected through observations, interviews, and record keeping forms are only as good as the observers, interviewers, and forms. Furthermore, naturalistic inquiry is not widely accepted by the scientific community because it is subjective, highly inferential, and it does not meet tests of rigor. Conversely, naturalistic inquiry is an extremely responsive form of evaluation and it adapts to a changing environment. The learning environment in developmental education is constantly changing; in fact, we strive for change. Furthermore, naturalistic inquiry is relatively cheap because it operates in tandem with what we do every day. And, for the most part, qualitative evaluation methods are simple and direct.

Utilizing naturalistic inquiry to evaluate learning will not be difficult for developmental educators because many programs already have established data collection systems and interview and observation procedures. Developmental educators need to put the data to work; and that is what naturalistic inquiry is all about.

SUMMARY

by Carol Clymer Walvekar

This booklet focused on the evaluation and assessment of learning in developmental education. The first section discussed the complexities of learning and the inherent difficulties in judging the quality and quantity of what is learned by students in developmental programs. It proposed a system for evaluating learning. This system was diagrammed and the result was a model which is generalizible to most developmental programs.

The major value of Section I was that it forced some structure on the rather indeterminate subject of evaluating learning. The major value of the next three sections was that they elaborated on specific methods for evaluating learning and thereby built upon the structure outlined in Section I. Section II provided much information on the types, uses, administration and value of tests for the assessment of learning. The message suggested in this section was that there are many tests available commercially for the developmental educator, they have shortcomings, but if used for the right reasons, they do an adequate job of assessing learning. Other tests, not available commercially, have to be developed locally. Section II gave some ideas for the development of teacher-made achievement tests while Section III discussed the use and development of criterion-referenced tests. Additionally, Section III demonstrated the necessity for using criterion-referenced tests in the developmental programs because of the high degree of individualization in most programs. It also offered very specific instructions for constructing these tests and using them in instructional sequences.

The final section of this publication showed how naturalistic methods of evaluation can be used to fill in the gaps that are left by tests. It defined naturalistic inquiry and described the basic principles for implementing it. Section IV also included several examples for using naturalistic inquiry to evaluate learning in developmental programs. Overall, this project attempted to supply practitioners with advice and guidelines for using and developing a variety of tests, and to provide suggestions for evaluating learning in areas where tests are not appropriate or adequate.

Typically, in other commentaries on testing, measurements and assessment, recommendations are made for more extensive research and development of tests and measurements that are more valid, reliable, consistent, objective and so forth. That recommendation will not be made here. Not because it has been said too often and not because it is not an accurate statement. We can always improve the tests and measurements that are available to evaluate learning.

Knowledge and learning are in constant states of flux and measurement instruments must adapt with and to these changes.

The recommendations that will be made are prequisite to developing new and better tests. First, in developmental education we must become thoroughly familiar with the tests that are already available. Reading and using this resource is a good start, but there are several other excellent resources which have been recommended here and they should be read and used as well. The subjects of evaluation and assessment are too vast to acquire expertise in a limited publication such as this. Further, every developmental program should seriously consider maintaining a test file of the most popular and common placement, diagnostic, and achievement tests. The Kellog Institute at Appalachian State University in Boone, North Carolina has printed a list of many tests used in developmental education. This list would be a good starting point for establishing the test file. If money is not available for a test file within the developmental program, it is possible that the director can go to other departments with common purposes such as counseling, education, or the library to cooperatively develop a test file. Once a file is established, developmental staff should become thoroughly familiar with the tests and the manuals.

Second, a point which has been made in each section of this booklet is that no one test, no one test type, can do the job of evaluating all the complexities of learning. Practitioners must use a variety of tests, adapt them to a variety of purposes, and realize that there will still be some unanswered questions about the learner. Thus, observations and interviews with learners are necessary. Use of a potpourri of assessment methodologies need not be viewed as haphazard or aimless if the tests are used creatively but wisely and sufficient supplemental information is obtained. This means that practitioners in developmental programs should try to implement an eclectic approach to evaluating learning.

Both recommendations made above can be time-consuming to implement. It is well-known that developmental educators do not have much time. Students usually want to complete developmental programs as quickly as possible, and there is frequently a lot to accomplish in a short amount of time. But it takes time to know where a learner is, where the learner is going, and how the learner is going to get there. For many developmental students, twelve years of schooling was not enough time to discover these things. Is it likely that 15 weeks or even one year will be enough time? The third and final recommendation is to carefully consider the amount of time that is taken to evaluate learning. Are students rushed in and out of placement tests and only the quantitative data on the tests analyzed? Are records kept on students' learning but barely looked at after they are written down? Are students bound to cognitive objectives determined through criterion-

referenced tests without consideration of affective influences on the learning? Is quality sacrificed for quantity for time's sake? Time is expendable, but students are not. We must take time to know the measurements we are using, use them carefully, and realize that even this much time may not be enough.

These recommendations present challenges for the developmental educator. I know they can be met. Challenge, in my opinion, is the core of developmental education in post-secondary institutions. We have many challenges everyday; what are a few more?

REFERENCES

The Introduction

Bruffee, K. A. "The Structure of Knowledge and the Future of Liberal Education." An unpublished paper, n.d.

Hilgard, E. R. & G. H. Bower. *Theories of Learning.* New York: Apple-Century-Crofts, 1966.

Lien, A. J. *Measurement and Evaluation of Learning.* Dubuque, Iowa: Wm. C. Brown, 1976.

Marx, M. H. & E. M. Bunch. *Fundamentals and Applications of Learning.* New York: MacMillan, 1977.

Section I

Ball, S. "Research and Program Evaluation: Some Principles, Precepts, and Practices." In *The Final Report of National Project II: Alternatives to the Revolving Door.* New York; Bronx Community College, n.d.

Wittrock, M.C. The Evaluation of Instruction: Cause and Effect Relations in Naturalistic Data," *Evaluation Comment,* May, 1969, Vol. 1, p. 1-7.

Section II

Ahmann, Stanley J. In *The Seventh Mental Measurements Yearbook.* O.K. Buros (Ed.). Highland Park, New Jersey: Gryphon Press, 1972, 9, 19-21.

Beatty, Wolcott, Ed. *Improving Educational Assessment and an Inventory of Measures of Affective Behavior.* Washington, D.C.: Association for Supervision and Curriculum Development, 1969.

Blanc, Robert and Michael Moninger. "An Interview with Robert Fuller." *Journal of Developmental and Remedial Education.* 5 (1), 5-9.

Boylan, Hunter. "Midterm Progress Report" Appalachian State University, Boone, North Carolina, Undated.

Breland, Hunter M. "Multiple-Choice Test Assesses Writing Ability." *Findings,* 1977, vol. 4 (1), 1-4.

Brown, Frederick, G. *Principles of Educational and Psychological Testing.* New York: Holt, Rinehart and Winstron, 1976.

Buros, O.K. (Ed.). *Tests in Print II.* Highland Park, New Jersey: Gryphon Press, 1974.

Cooper, Charles. *The Nature and Measurement of Competency in English.* Urbana, Illinois: The National Council of Teachers of English, 1981.

Eagle, Norman. In *The Eighth Mental Measurements Yearbook.* O.K. Buros (Ed.). Highland Park, New Jersey: Gryphon Press, 1978, 475, 641-643.

Ebel, Robert, L. "Educational Tests: Valid? Biased? Useful?" *Phi Delta Kappan.* October 1975, 83-88.

Findley, Warren G. In *The Eighth Mental Measurements Yearbook.* O.K. Buros (Ed.). Highland Park, New Jersey: Gryphon Press, 1978, 12, 40-43.

Ivens, Stephen. Telephone Interview on December 18, 1981. New York City Office of College Boards.

Lappan, Glenda. In *The Eight Mental Measurements Yearbook.* O.K. Buros (Ed.). Highland Park, New Jersey: Gryphon Press, 1978, 292, 436-437.

Lauridsen, Kurt. Plenary Session on Learning Center Issues. The Non-Traditional Student in the College Learning Center. Long Island University, Brooklyn, N.Y. May 14-16, 1981.

Maxwell, Martha. *Improving Student Learning Skills.* San Francisco: Jossey-Bass, 1980.

Popham, W. James. *Educational Evaluation. Englewood Cliff, New Jersey: Printice-Hall, Inc., 1975.*

Purves, Alan. "Competency in Reading." In The Nature and Measurement of Competency in English, Charles R. Cooper (Ed.). Urbana, Illinois: National Council of Teachers of English. 1980, 65-91.

Roark, Albert E. and Scott A. Harrington. In *The Seventh Mental Measurements Yearbook*. O.K. Buros, (Ed.). Highland Park, New Jersey: Gryphon Press, 1972, 782, 1211.

Roekel, Byron H. In *The Eighth Mental Measurements Yearbook*. O.K. Buros (Ed.). Highland Park, New Jersey: Gryphon Press, 1978, 777, 1298-1299.

Rosenbach, John. In *The Eighth Mental Measurements Yearbook*. O.K. Buros (Ed.). Highland Park, New Jersey: Gryphon Press, 1978, 28, 91-94.

Schultz, Richard E. In The Eighth Mental Measurements Yearbook. O.K. Buros (Ed.). Highland Park, New Jersey: Gryphon Press, 1978, 28, 95-96.

Shay, Carlton, B. In *The Seventh Mental Measurements Yearbook*. O.K. Buros (Ed.). Highland Park, New Jersey: Grypon Press, 1972, 782, 1210-1211.

Smith, Fred M. and Sam Adams. *Educational Measurement for the Classroom Teacher*. New York: Harper and Row. 1966.

Sowder, Larry. In *The Eighth Mental Measurements Yearbook*. O.K. Buros (Ed.). Highland Park, New Jersey: Gryphon Press, 1978, 292, 437-438.

Stanley, Julian C. and Kenneth D. Hopkins. *Educational and Psychological Measurement and Evaluation*. Englewood Cliffs, New Jersey: Prentice-Hall, 1972.

Thorndike, Robert L. and Elizabeth P. Hagen. *Measurement and Evaluation of Psychology and Education*. New York: John Wiley and Sons, 1977.

Wallace, Gerald and Stephen C. Larsen. *Educational Assessment of Learning Problems: Testing for Teaching*. Boston: Allyn and Bacon, Inc., 1978.

Section III

Block, J. H. "Introduction to master, learning: Theory and practice," in J. H. Block (ed.), *Mastery Learning: Theory and Practice*. New York: Holt, Rinehart & Winston, 1971.

Bloom, B.S. "Mastery Learning," in J.H. Block (ed.), *Mastery Learning: Theory and Practice*. New York: Holt, Rinehart & Winston, 1971.

Cartier, F. A. "Criterion-referenced testing of language skills." *TESOL Quarterly*. February, 1968, pp. 27-32.

Gilman, D. A. "Alternatives to testing: Criterion-referenced measurement," in *Alternatives to Tests, Marks, and Class Rank*. Terre Haute: Indiana State University, 1974.

Glaser, R. "Instructional technology and the measurement of learning outcomes," *American Psychologist*. September, 1963, pp. 519-521.

Greco, T. H. "Is there really a difference between criterion-referenced and norm-referenced measurement?" *Educational Technology*. August, 1974, pp. 22-25.

Gronlund, N.E. *Preparing Criterion-Referenced Tests for Classroom Instruction*. New York, Macmillan, 1973.

Haladyna, T. "The paradox of criterion-referenced measurement," Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, April, 1976. (ERIC ED 126 155)

Hambleton, R. K. "On the use of cut-off scores with criterion-referenced test in instructional settings," *Journal of Educational Measurement*. August, 1978, pp. 277-290.

Popham, W. J. *Criterion-Referenced Measurement*. Englewood Cliffs, N.J.: Prentice-Hall, 1978.

Popham, W. J. & Husek, T. R., "Implications of criterion-referenced measurement," *Journal of Educational Measurement*. June, 1969, pp. 1-9.

Roberson, D. R. *Development and Use of Criterion-Referenced Tests*. Austin, TX: Educational Systems Associates, 1975.

Smith, C. W. "Criterion-referenced assessment in contrast to norm-referenced measurement," *Marquette University Educational Review*. April, 1973, pp. 23-32.

Tuinman, J. J. "Criterion-referenced measurement in a norm-referenced context," in Samuels, S. J. (ed.), *What Research Has to Say About Reading Instruction.* Newark, DE: International Reading Association, 1978.

Section IV

Guba, E. G. & S. L. Lincoln, *Effective Evaluation.* San Francisco: Jossey Bass, 1981.

Patton, M. G. Qualitative Evaluation Methods. Beverly Hills: Sage, 1980.

ABOUT THE AUTHORS

Carol Clymer Walvekar is the Assistant Director of Study Skills and Tutorial Services at the University of Texas at El Paso. She worked as the Coordinator of Services for the Center for Learning Assistance for five years. Walvekar is currently the Chairperson for the Research and Evaluation Committee of NARDSPE and the Task Force for Accountability and Evaluation for ACPA, Commission XVI on Learning Centers. Additionally, she edits a column on Evaluation for the WCRA Newsletter.

Nancyanne Rabianski is the Director of the Developmental Studies Program/Learning Assistance Center at the University of New Haven. Her Ph.D. is in Research and Evaluation (SUNY at Buffalo, 1979). She has published journal articles and spoken at conferences regarding test evaluation, learning styles, and writing anxiety. She is a member of NARDSPE and has attended the Kellogg Institute for Developmental Educators.

Dr. Karen G. Smith is the designer, developer and director of the Center for Learning Assistance at New Mexico State University, where she received her M.A.T. and Ed.D. degrees in reading education and educational management. Her doctoral dissertation, "The Development and Analysis of a Criterion-Referenced Test of Textbook Reading and Vocabulary Discrimination for University Freshmen with Reference to Their ACT Composite Scores," was published in 1980. She has supervised the development of criterion-referenced tests for use in the basic skills areas and competency demonstration assessment for study skills improvement in the Center for Learning Assistance.

ERIC Clearinghouse for
Junior Colleges
NOV 7 1986